DOCUMENT RESUME

ED 239 004 UD 023 307

AUTHOR Cook, Thomas D.
TITLE What Have Black Children Gained Academically from School Desegregation: Examination of the Metaanalytic Evidence.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 31 Aug 83
NOTE 61p.; For related documents, see UD 023 302-308. Paper submitted as one of a collection from the National Institute of Education Panel on the Effects of School Desegregation.
PUB TYPE Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; *Black Students; *Desegregation Effects; *Effect Size; Elementary Secondary Education; *Mathematics Achievement; Meta Analysis; Outcomes of Education; Program Effectiveness; Program Evaluation; *Reading Achievement; Research Reports; School Desegregation

ABSTRACT

This paper analyzes the 19 studies presented to the National Institute of Education's (NIE) panel on the effects of school desegregation on black achievement and discusses the author's own findings. The author concludes that desegregation did not cause any decrease in black achievement generally, nor did it cause any increase in math achievement. Although desegregation increased mean reading levels, the distribution of reading effects appeared to be skewed, with a disproportionate number of school districts obtaining atypically high gains. Studies with the largest gains were characterized along a number of methodological and substantive dimensions (none of which could be isolated as causes of the atypically high reading gains) including: small sample size, two or more years of desegregation, desegregated children who outperformed their segregated counterparts even before desegregation began, and desegregation that occurred earlier, was voluntary, occurred in schools with larger percentages of whites, and was associated with enrichment programs. Because of the small samples in the NIE project, and the apparently non-normal distributions, the author states he is not confident that anything has been learned about desegregation's effects on reading on the average. Across the few studies examined, he found that variability in effect sizes was more striking and less well understood than any measure of central tendency. The paper ends with a review of the implications of the findings for various interest groups and a summary of the implications the NIE project has for theories of research synthesis. (CMG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED239004

What Have Black Children Gained Academically from School Desegregation: Examination of the Metaanalytic Evidence

Thomas D. Cook
Northwestern University

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Final Draft
August 31, 1983

Paper submitted as one of a collection from the National Institute of Education Panel on the Effects of School Desegregation.

ED239004

INTRODUCTION

My assignment is to comment on the foregoing essays by Armor, Crain, Miller, Stephan, Walberg and Wortman in order to help readers decide what should be concluded from prior evaluations of how school desegregation has affected the academic achievement of black children. All but two of the essays contain a metaanalysis by the author. Crain's paper is one of the exceptions. Instead of conducting a metaanalysis, he critically discusses some of the assumptions behind the others' efforts and concludes that he will stand by the results of his own prior metaanalytic work (Crain & Mahard, 1983). I shall refer to his prior metaanalysis based on 93 studies more than to his essay in this volume. Walberg is the other exception. He devotes most of his essay to a review of factors other than desegregation that raise academic achievement. He does this to make the point that, if the purpose of desegregation is to raise the achievement of black children, then more effective means exist to do this than desegregation. Walberg does however, reanalyze three prior metaanalyses--by Krol (1975), Crain & Mahard (1982) and Wortman, King and Bryant (1982)--in order to make the further point that, in his estimation, the average effect sizes they present do not reliably differ from zero. I intend to deal with his statistical analysis to a small extent, but will not deal directly with his larger point about relative efficacy.

The first part of the present paper deals with the metaanalytic work of Armor, Miller, Stephan and Wortman, and is largely restricted to the 19 studies

selected by the panel. The purpose is to arrive at an estimate for this sample of how desegregation has affected the achievement of black children. I try to restrict my commentary to the most important points and assumptions made by the authors, and make no attempt at a comprehensive analysis of any single person's work in order to be comprehensive about its strengths and weaknesses. This is to keep the focus on the desegregation issue. In the second part of the paper I take my own results, which are both similar to and different from those of the panel, and discuss several ways they can be interpreted. In particular, I ask how generalizable are results from the panel's 19 studies when they are compared to the results from larger data bases; I probe the extent to which my findings speak to the information needs of groups with different stakes in school desegregation; and I speculate about whose interests the panel's results might advance or prejudice.

RESULTS

1. <u>The Studies Examined</u>. Individual panel members considered different subsets of the 19 studies that most of them deemed methodologically adequate. Armor dropped the study by Rentsch on grounds, first, that the desegregated group and the segregated controls differed by so much initially; second, that the pretests and posttests involved different measures; and third, that the desegregated control group contained some white children. He also dropped the study by Thompson & Smidchens on grounds that the segregated controls were in classes made up only 42% of minority students. However, he included the study by Carrigan, even though its segregated control group members were in classes that

were hardly more "segregated"--50% minority. Indeed, Miller and Stephan dropped the Carrigan study because of its questionable control group. In a few other cases, Armor selected control groups within a study that differed from the choice of all other panelists. The net result of Armor's preferences was lower effect sizes since (1) Rentsch obtained some of the largest effect sizes, (2) Carrigan resulted in both positive and negative effect sizes, and (3) both Rentsch and Carrigan involved multiple comparisons and so their results were disproportionately weighted whenever comparisons were the unit of analysis rather than individual studies.

Miller dropped both Carrigan and Thompson & Smidchens from his analyses because the segregated controls were not segregated. He also differed from the other analysts in preferring to compute an effect size per study instead of per comparison. Much has been written in the metaanalysis literature on this topic, and our preference is to compute or report effect sizes each way. However, if only one choice is available, we favor a sample of studies because this does not weight the results in favor of school districts where desegregation was tested using several grades.

Stephan also omitted the studies by Carrigan and by Thompson & Smidchens. However, he also objected to the studies by Iwanicki & Gable and by Slone on grounds that they dealt with the second year of desegregation while other studies dealt with the first year. He further objected to Slone because the segregated controls were attending a school that was 40% white. This left Stephan with only 15 studies to analyze. Since the studies he omitted all tended, with the exception of Slone, to have zero or negative effect size estimates, it is clear that Stephan's sampling decision disposed his analysis

towards a larger average effect size than other panelists.

Wortman differed from the other panelists in two important ways. First, he preferred his own selection of 31 "superior" studies to the panel's 19. However, his analysis of the 31 showed that designs without control groups produced higher effects size estimates than designs with control groups. Hence, I treat his analyses based on studies with controls differently from the analyses without controls for, among other possible artifacts, maturation and testing effects can inflate estimates of the desegregation effect. Second, in his analyses of the panel's 19 studies, Wortman was more strict than the others about what he would accept as valid information about variances. Since such information is crucial for computing effect sizes he was able to produce estimates that also controlled for pretest differences between the desegregated and segregated control groups for only 11 of the 19 studies favored by the panel. One of these was the study by Carrigan. Omitted were Clark, Evans, Iwanicki & Gable, Klein, Laird & Weeks, Slone, Syracuse, and Thompson & Smidchens. Since Wortman preferred somewhat different standards of methodological adequacy than the panel, I sometimes include estimates computed from his analyses of the 11 panel studies, and at other times estimates based on the larger subset of his preferred studies that involved designs with control groups. [redacted]. These studies should overlap heavily with the panel's selection criteria.

The panelists provided estimates for reading and math combined, for reading alone, and for math alone. It is interesting to note that there is no obvious relationship between gains in mathematics and reading when the desegregated are compared to the segregated. To compute a correlation of reading and math gains

would not be useful because of the small number of studies and comparisons for which there were measures of both reading and mathematics gains. However, of Armor's 18 relevant comparisons, math and reading gains had the same sign in seven instances, different signs in eight, and three instances were indeterminate because of zeros. Of Miller's 13 comparisons, seven had the same sign and six the opposite; while of Stephan's comparisons there were 13 with the same sign, 11 with the opposite, and one was indeterminate. Math and reading gains were not clearly related, and little is gained by adding them together. Consequently, I prefer to present results separately for each knowledge domain. However, for purposes of continuity with the panelists some of my reanalyses will involve reading and math scores combined. When that happens, my analyses--like those of the panelists--weight reading slightly more than math because more reports included reading than math measures.

2. Panelists' Results. Using his own preferred set of studies based on a sample of comparisons. Armor obtained an effect size of .06 for reading and .01 for math; Miller obtained an effect size of .16 for reading and .08 for math; Stephan's values were .15 and .00; while in my analysis of Wortman's results for the eleven studies with pretest adjustments, the mean effects were .26 and .08. (Wortman's own results from the panel's 19 studies were .28 and .23, but this includes studies where no pretest adjustments were made. His estimates from his total sample of 31 studies were .57 and .33, but these are based on some studies without control groups. Thus, I consider both of these last sets of estimates to be problematic).

If we turn now to estimates of reading and math combined, Armor's overall

estimate was .04, Stephan's was .14 (but .07 when computed as gain per 8 month school year), Miller's was .12, while Wortman's was .17 derived from the studies of his own choosing that had control groups.

If one took the panel's estimates at face value they would appear to support the following conclusions:

1. Desegregation did not cause a decrease in the achievement of black children.

2. It probably did not cause an increase in math skills, for the mean gains vary from 0 to .08 standard deviation units.

3. It may have caused an increase in reading skills, for the mean gains vary from .06 to .26.

The range estimate for reading deserves comment, since the upper bound comes from our analysis of Wortman's eleven studies where pretest adjustments could be made. This is a considerably smaller sample than the other authors analyzed, and so should be treated as particularly tentative. Omitting it gives a revised range that permits a fourth conclusion, which I believe to be better justified than the third conclusion immediately above:

4. The gain in reading was somewhere between .06 and .16 standard deviation units. This is between two and six weeks of gain if we follow the rule of thumb of Glass et al (1981) and associate a gain of one-tenth of a standard deviation with one month's gain in knowledge.

The small discrepancies between the panelists in mean estimates principally reflects differences in (1) the studies included for review; (2) the way effect sizes were computed; and (3) a preference for some types of control groups over others within a few studies. I shall resist the temptation to discuss each of

these issues in order to make judgments for each of them about the methodological option to be preferred, after which point estimates of gains could be computed. While such an exercise would result in easily remembered single number estimates of reading and math gain, the resulting precision would be misplaced. In metaanalysis, varying the assumptions underlying an analysis is desirable because it makes heterogeneous those facets of research where no "right" answer is available and fallible human judgment is required. To attempt to legislate a single "right" way either to compute effect sizes or to sample studies would be counterproductive so long as none of the analysts is clearly wrong. Indeed, the idea of selecting a panel of methodologically sophistocated experts with different views on school desegregation is predicated on the particular utility that would result if the panel's estimates of desegregation's effects converged <u>despite the differences in values and methodological predilections of individual panelists</u>. It is more reasonable to expect "convergence" as a range than a point. To search for the elusive "true" point estimate of effect could involve laborious debates about fine points of methodology and substance that might occur within a range of estimates that many would think has few practical implications.

Speaking personally, I am impressed by the degree of correspondence between the panelists when only the 19 core studies are considered. None achieves negative estimates; all achieve larger estimates for reading than math; and the largest single difference—between Armor and Miller for reading gains—is of a magnitude many would consider small—viz., a difference of about one month of gain.

The convergence is all the more dramatic since, across all dependent

variables, Krol obtained an estimate of .10 from his own metaanalysis of "better" desegregation studies, while a similar estimate resulted from Crain & Mahard (1983) when one aggregates across all their dependent variables for the randomized experiments and studies with both pretest-posttest measurement and control groups of segregated black children. Combining math and reading and analyzing only the studies preferred by the present panelists, Armor's estimate was .04, Miller's was .12, and Wortman's was .17 for all the studies he found with pretests and black control groups, while Stephan's estimate was .14 without his correction for the length of time desegregation had been taking place--a correction that none of the other panelists made. The average of the panelists values is .11, only slightly higher than the estimate obtained by Krol and Crain & Mahard. (However, as we later see, Crain rejects this estimate, preferring to base his judgment on studies where desegregation occurs at kindergarten or first grade.)

3. The Distribution Problem. As a measure of central tendency the mean depends on a normal distribution of scores. In Figures 1 through 4 we present frequency distributions of reading effect sizes for Armor, Miller, Stephan, and Wortman based on the studies they chose to analyze. (For Wortman we add the math data since he presents reading effect sizes for only eleven studies where pretest adjustments were made, and this results in a particularly poor estimate of the distribution). In all cases except Miller the sample sizes are based on comparisons rather than studies. But irrespective of the unit of analysis, the distributions are visibly skewed, with a disproportionate number of effect sizes falling in the upper range.

Table 1 presents the medians and modes corresponding to the reading mean.

The median is computed for a sample of both comparisons and studies and is defined as the value of the (N+1)/2th case. To compute a mode with so few cases, we constructed a scale composed of categories with intervals of .10 standard deviation units whose midpoints are presented in Figures 1-4. Each effect size was assigned to its respective category, with scores of zero being assigned in equal proportions to the category 0 to +.10 and 0 to -.10. For Miller, no value is reported for the median of comparisons since he only provided data on studies. Sometimes, no mode is presented for Wortman because his smaller sample of studies from the panel's set that had pretest adjustments often makes it difficult to determine any modal category with more than three cases falling into it.

Table 1 shows that mean effect sizes for reading are larger than median effect sizes irrespective of whether the latter are computed as a median of comparisons or of studies. It also shows that the mode is smaller than the other measures of central tendency and hovers around zero. Indeed, the mean of the mean effect sizes across all four panelists is .15, the mean median of comparisons is .08, the mean median of studies is .05, while the modal categories are of effects between +.05 and -.05!

Insert Figures 1 through 4 about here

Table 1 was recomputed based on the 17 core studies most panelists agreed upon. That is, Thompson & Smidchens was omitted since three of the four panelists who did metaanalyses questioned it; and Carrigan was omitted since at least two of the panelists objected to the questionable nature of their "segregated" controls. In computing the data for Armor, the missing values for

Rentsch were taken from Wortman. Stephan provided his own estimates for the studies by Iwanicki & Gable and Slone that he preferred to leave out of most of his own analyses. As Table 2 shows, having a common set of studies reduced the dispersion of mean effect sizes for reading. The range for the panelists—Wortman excepted because his analysis is not based on the 17 studies and I did not want to take his six missing estimates from other panelists since that would involve estimating about 30% of the scores--the range shifted from .06--.16 to .13--.16. However, even with the same 17 studies per analyst the table still shows that medians are lower than means, and that modes are lower than medians.

Insert Tables 1 and 2 about here

A corresponding table for math from the authors' own preferred set of studies is in Table 3. Modes could not reasonably be computed due to the smaller number of math than reading comparisons. However, the means are consistently higher than the medians.

Combining math and reading allows modes to be computed again and results in the same basic relationship between measures of central tendency. This is true whether one uses the authors' own set of preferred studies (Table 4) or the common set of 17 (Table 5). The individually preferred studies produce a range of mean estimates from .06 to .16, of median estimates from .00 to .08, and of mode estimates from -.15 to +.05.

Insert Tables 3, 4, and 5 about here

These differences in central tendency result because the distribution of effect sizes is skewed. The skewness means that, if one were willing to assume that the present results are applicable to the nation at large today--a dangerous assumption!--then (1) for any school district that desegregates the most reasonable expectation is that there will be no effects on black achievement, for the mode suggests that this outcome is obtained more often than any other; (2) 50% of the school districts will probably raise achievement by about three-one hundredths of a standard deviation (the average median of studies across the panelists), while 50% of them will probably raise it by less than this; but (3) the national impact will be to raise the achievement of black children in reading by between two and six weeks and to raise achievement in math, if at all, then by something less than three weeks—the upper range of mean estimates. However, (4) a minority of school districts could expect to make larger positive gains. Using Miller's reading estimates for the moment, larger gains appear to have been obtained by Anderson (.733), Beker (.400), Syracuse (.691), and Zdep (.671). In mathematics, the outliers were less common but still visible (Anderson, .669, Klein .333, and Van Every .543).

But Stephan's estimates make the studies with outlying results seem less extreme and some different outliers emerge. He computes effect sizes in a way that controls for the length of time children have been under study in a desegregated school. When reading effect sizes are computed <u>per eight month school year</u>, the outliers are pulled in because they tended to come from studies lasting two or three years. The new values are: Anderson (.42), Beker (.13),

and Zdep (.66). (Stephan leaves Syracuse out of his sample). For mathematics, the positive outliers now become: Anderson (.24), Klein (.33), and Van Every (.14). Stephan's computation of effect sizes leads to less variable and less skewed estimates than the other panelists, which is why medians and modes make less of a difference to his computations of central tendency than to others. But the choice of a measure of central tendency still makes a difference in Stephan's estimates, for both reading and reading and math combined.

However, Stephan's work does present a puzzle. He is the sole panelist to compute a median, and on page 24 of his report he mentions that the median gain in verbal achievement (reading) is .13. (His corresponding means were .17 for the sample of comparisons and .15 for the sample of studies.) I have examined Stephan's effect sizes from his Table 1 and have been unable to arrive at the same value. My own estimate based on a sample of comparisons and omitting the studies he leaves out is .08. Readers should scrutinize Stephan's Table 1 and estimate for themselves the effect size for reading scores above which 50% of the effect sizes fall and below which 50% fall.

4. <u>The Confidence Problem</u>. Our reanalysis of the panelists' studies using multiple measures of central tendency should not be intepreted to mean that, in our opinion, desegregation has had no effect on most schools. There are two reasons for a low level of confidence in the results presented in Tables 1 through 5. First, we do not know the underlying distribution of mean effect sizes (however computed) for the population of school districts that have already desegregated. It is not clear how representative the panel's core set of studies are. Second, with so few comparisons and studies, we cannot have

much confidence in the sample distributions presented in Figures 1-4. A dozen new cases could radically alter each of the estimates of central tendency. With such a poorly estimated and unstable distribution it is not clear that the mean would remain unchanged, even if more cases were added from the very same population that the present sample is supposed to represent.

Statistical significance tests are typically used to make inferences about the level of confidence one should ascribe to findings. (Because of lay misunderstandings of the word "significance", we prefer to talk of tests of statistical reliability rather than statistical significance.) Walberg has maintained that for measures of math and reading combined, none of the estimates obtained by Krol, Crain & Mahard and Wortman, King & Bryant reliably differ from zero. In the current case, our calculations of reliability indicate that: (1) For Armor, the mean estimates for math alone and for reading and math combined do not differ from zero, but the estimate for reading does so marginally ($p<.10$); (2) for Miller, the estimate for math does not reliably differ from zero, but the estimates for reading alone and for reading and math combined do so; (3) For Stephan, the effect for math is not reliable, while for reading and for math and reading combined, conventional levels of statistical reliability are reached irrespective of whether the mean is computed with or without correction for the length of desegregation; (4) For Wortman, the effects for reading and for reading and math combined both differ from zero even when we consider only the small sample of studies with pretest adjustments.

These statistical tests are themselves partly problematic. In all cases except Miller, the analyses are based on a sample of comparisons. But since some studies produce more than one estimate of effect size, the assumption of

independent errors may not be met. This particular problem does not occur in Miller's analysis. But there the small sample of studies increases the dependence on the assumption of a normal distribution of effect sizes. But as the difference between the various measures of central tendency indicates, the distribution of effect sizes may not be normal. Hence, all the statistical test results reported above (and in Walberg) should be treated with some caution. As they stand, they suggest that neither the mean reading effect nor the mean effect for reading and math combined is due to chance.

However, to complicate matters it is not likely that the medians and modes differ from zero. The standard error of a median is normally set at 125% of the value of the standard error of the means from the same distribution, reflecting the greater instability of medians. By this criterion, no medians reliably differ from zero for reading or for reading and math combined. No estimate of the reliability of modes is necessary since they hover so closely around zero. However, the medians and modes are based on so few cases that estimates could shift radically once a dozen new values are added to the distribution.

If the population of effect sizes is indeed skewed, it is not clear which measure of central tendency is to be preferred. The mean represents national impact at some abstract, aggregate level, and is of use to those persons and groups most interested in gaining a national perspective on education and society. The mode represents what should happen to the typical school, and so may be of most interest to any school district or judge considering desegregation, especially if the district in question differs from those where desegregation has produced large impacts in the past—characteristics we shall explore below. For any commentator willing to assume that the distribution of

effect sizes in the population approximates the (unclear) sample distributions we have obtained, it is important to decide at a high level of consciousness on the different utilities implicit in different measures of central tendency.

5. Why do Some School Districts Show Larger Gains in Reading? The skewness in the distributions indicates, not only that the mean may be a misleading measure of central tendency, but also that it might be productive to probe the reasons why some school districts are outliers. Discovering what they did to achieve larger gains could, for instance, be used to develop specific guidelines for desegregation plans, which school districts could then select if they believed they were suitable for their schools. But since desegregation is an amorphous set of activities that differs from site to site, and since we have so few studies, no one should expect a definitive answer to the question of what characterizes school districts with large reading gains. At most, one should expect grounded hypotheses to emerge. Our discussion is in two parts: Which were the districts with large gains; and what differentiates them from other districts?

(a) Which Were the School Districts with Larger Reading Gains? Before probing substantive reasons for high reading gains, it is important to raise three methodological issues that reduce confidence in judgments about the identification of valid outliers. The sample sizes in the studies under review vary considerably, from 12 desegregated children in Zdep to over 1,000 in Sheehan and Marcus. Several panelists analyzed the relationship between sample size and effect size, concluding that smaller samples tended to produce larger

estimates but that the relationship was not reliably different from zero. Considering classical sampling theory in isolation, we would not expect sample size to be linearly related to effect sizes without transformation of the original metrics. In a normal distribution with mean equal to zero, we would expect smaller samples to produce larger estimates, but in equal proportions each side of zero. This is equivalent to a negatively accelerated decay function when plotting effect size against sample size, irrespective of the sign of the effect. Figure 5 presents the mean reading effect size, free of sign, for studies with desegregated samples of 20 or less, between 21 and 30, between 31 and 40, 41 and 50, between 50 and 100, and over 100. An overall relationship is apparent that might well be of the expected quadratic form, though with such a small sample of studies it is hard to be sure. More important, though, is that with such a small sample of studies it is possible for more of the studies with smaller samples to fall on one side of the mean than the other. If we take the studies identified from Miller's estimates as outliers, we note the following individual sample sizes in the desegregated groups for analyses of reading: Anderson (34), Beker (36), Syracuse (24), and Zdep (12). This is a total of 106 desegregated children. Since a total of 2812 were studied for reading, the outliers responsible for the higher mean estimates constitute about 4% of the total sample of desegregated children, but are about 25% of the studies Miller analyzed (4 of 17). If we add Rentsch to the list of outliers because analysts other than Miller and Stephan place him there, then the outliers represent 30% of the schools studied (5 of 17) but only 7% of the children!

Insert Figure 6 about here

A second methodological reason for caution in substantively pursuing why some school districts have large gains is also related to sampling instability. If we were to define positive outliers in terms of their gains in both reading and math, few of the outliers would be the same as when reading was considered alone. Thus, the unweighted gain in Anderson, using Miller's estimates, was .70, for Beker was .19, and was .26 for Zdep. (It was .035 for Rentsch in Miller's analysis). When a joint criterion is used to define outliers, only Anderson clearly emerged. Indeed, the three other studies had negative estimates for math! Pursuing the instability theme further leads us to note that the second largest negative outlier for reading (Van Every, -.17) is based on a desegregated sample of only 20, and the math estimate is +.54! We are not arguing that desegregation should have affected both reading and math. We are only suggesting that we would be more confident of having identified valid outliers if reading and math gains were correlated among the potential outliers.

The third methodological issue concerns how effect sizes were computed. All the panelists are commendably sensitive to the need to control for differential growth rates between the nonequivalent desegregated and segregated control groups, and all go about the task in similar--but not quite identical--ways. The adequacy of statistical adjustments for selection-maturation depends on many factors, including the (unknown) true selection difference, the reliability of measures, the comparability of within-group regression lines, etc. In metaanalysis, the hope is that, across all the studies examined, the inevitable imperfections in the analysis of any

one study will even out so that the average bias due to selection-maturation will be zero. However, there is no presumption that the bias will be zero in any single study. Yet in analyzing outlier effect sizes one has to assume that the average selection and selection-maturation bias among the outliers is zero. However, one might easily have capitalized on chance and have isolated the subset where adjustment has been the least adequate. Indeed, in four of the five outlier cases the desegregated children outperformed the segregated initially, and in the other case the means were essentially identical.

Thus, the possibility cannot be ruled out that the outliers reflect: (1) sampling instability due to small sample sizes; (2) sampling instability that makes high reading gains not synonymous with general achievement gains; and (3) an underadjustment for initial group differences in reading achievement. It is within the limitations afforded by these three points that I now examine substantive characteristics of the outliers for reading.

(b) The Characteristics of Outlier School Districts. As previously discussed, one characteristic of the outlier school districts on Miller's list is that they evaluated longer periods of desegregation—up to three years in some cases. The relationship between effect sizes and length of desegregation is not clear due to sampling instability, with all the panelists who tackled the issue concluding that effect sizes seem larger in the five studies with two years of desegregation than in the nine studies with one year of desegregation. However, estimates seem to be lowest of all in the three studies with three years of desegregation! Since two year studies predominate among the studies with larger effects in Miller's Table 2, is suggests that effect sizes may be related to the

amount of desegregation that has taken place.

The predominance of two year studies among the districts with larger effects also leads me to prefer Stephan's estimates for defining outlier school districts. But to use his data I averaged his estimates across grades to give a single reading mean per study. The outliers fall into two groups: Anderson (.49), Syracuse (.58) and Zdep (.66) are in the one and Klein (.23), and Rentsch (.22) in the other. Even listing these outliers raises once again the spectre of instability, since Klein would not be an outlier for Miller, while Beker would be for Miller but not for Stephan!

Two substantive factors are associated with Stephan's larger effect sizes. One concerns when desegregation takes place. Figure 6 shows effects sizes per eight months of desegregation plotted against when desegregation began. The latter values are taken from Wortman rather than Stephan, since the information about grades in Stephan's Table 1 appears to be based on the grade at which desegregation began in some cases and on the grade when it ended in others. Figure 6 shows a clear negatively accelerated decay curve, with larger effects the earlier the desegregation. None of the panelists obtained effects of grade on achievement that were as clearcut as this, probably because they computed linear relationships, truncated at inappropriate grade levels, did not adjust effect sizes for the length of desegregation, or they assessed the grade of children when the study ended. Figure 6 suggests that at second grade a gain is obtained of about .30 standard deviation units per eight month year--though this estimate is based on only four studies!--that at the third grade the gain is .12 (five studies), while it is .14 at the fourth grade (based on nine studies).

In trying to explain why a small set of school district produced large

reading gains that skewed the distribution of effect sizes, it is important to probe whether the desegregation was voluntary or mandatory. According to Crain's report in this volume, all of the school districts I have identified as positive outliers had voluntary programs. This is perhaps not surprising, since the programs were voluntary in 15 of our 19 studies. For reading, only three school districts showed overall negative effects in Stephan's analysis--Sheehan & Marcus (-.07), Smith (-.01) and Van Every (-.12). The first and last of these were mandatory programs. Of the two other mandatory programs in the panel's sample, the study by Carrigan was omitted from some analyses but, when aggregated across grades, it produced a small negative effect. The other mandatory study produced a trivial gain of .02 across grades (Evans). It is clear, then, that mandatory programs were not associated with reading gains but that voluntary programs were.

However, the relationship between effect size and the voluntary/mandatory nature of desegregation could only be considered causal for these four cases of mandatory desegregation if all other interpretations of the relationship could be ruled out. However, two of the studies--Evans and Sheehan & Marcus--were done in Texas, were the only ones to use the Iowa Test of Basic Skills, and were two of the only three studies of desegregation activities that began in the 1970's. (The other study with apparent negative outcomes--Van Every--took place in Flint, Michigan, began in 1969, used the SRA test, and had very small samples.)

Just as it would be wrong to conclude with confidence that mandatory programs produce no gains in reading, so it would be wrong to conclude from the panel's core studies that desegregation beginning in the earlier grades results

in larger positive gains. There are signs of each relationship, but with only four mandatory programs and four second grade samples it is inevitable that we have not made heterogeneous all the sources of irrelevancy that might have produced spurious results. The reality is that if the sample size of studies is too small to permit a meaningful analysis of central tendency across 19 studies, it is even less appropriate for conducting responsible internal analyses to try to explain why some school districts seem to have achieved larger effect sizes than others.

This is true, not only of the potential explanatory factors analyzed above, but also of other factors about which individual panelists have speculated. Stephan points out that studies conducted at an earlier date tend to show larger effects, while Miller suggests that school districts with larger effects may have introduced enrichment programs at the time desegregation occurred and may have had smaller percentages of blacks in the desegregated classrooms. With the small samples on hand, it is inevitable, first, that no strong probes of the impact of such moderator variables is possible; and second that many interpretations remain to explain why some districts achieved particularly large positive or negative gains.

The points we want to stress are that: (1) the form of the distribution of effect sizes is not clear either for the population of school districts that have desegregated or even for the small sample of districts we have analyzed; (2) there may be districts that benefitted more from desegregation than other districts—but if so, it is not clear whether they are outliers for irrelevant methodological reasons (small sample sizes, unstable measures, or initial group achievement differences not completely adjusted away) or for relevant

substantive reasons; and (3) of the relevant substantive reasons, several are contenders as explanatory constructs but their unique contrubution cannot be unconfounded from the contribution of other factors. The factors at issue include: the child's grade at desegregation, the number of years of desegregation, whether the desegregation is voluntary or mandatory, the percentage of whites in the class, the copresence of desegregation and new enrichment programs, and the year in which desegregation took place.

6. Summary of the Reanalyses. A casual reading of the panelists' papers leads to the four conclusions mentioned earlier that are based upon the panel's 19 studies and seem quite consonant with the findings of prior metaanalyses by Krol and by Crain & Mahard that involved larger samples. These conclusions are: (1) desegregation does not decrease the achievement of black children; (2) it probably does not increase math achievement; (3) it probably raises reading scores; and (4) the increase in reading scores is somewhere between .06 and .16 standard deviation units or about two and six weeks. These last estimates were computed from 17 studies, about half of which dealt with a single year of schooling, and then usually the first one after formal desegregation began.

Our own analyses corroborate the first two of these findings. We continue to find no evidence that desegregation decreases achievement or that it increases achievement in math. Our differences involve the conclusions about reading. The present analysis suggests that whether there is an effect or not depends on the measure of central tendency used, with statistically reliable results emerging for mean gains but not for median or modal gains. The implication of the lower medians and modes is that the mean differences are

found, not so much because the "average" effect of desegregation on reading is positive but because--in the panel's sample at least--some school districts made atypically large reading gains that skewed the distribution of effect sizes.

It is therefore difficult to make an estimate of the size of the reading effect. There is one range estimate for the mean (between .13 to .16 when the same 17 studies from the panel's 19 are used with each analyst's own effect size computations--see Table 2), another range estimate for the median (.00 to .08 irrespective of the samples used--see Table 1 or 2) and yet another for the modal effect (between -.05 and +.05--see Tables 1 and 2). Combining the reading and math effect sizes makes no difference to the conclusion that central tendency values differ. The estimated means vary between .07 and .16 for the 17 common studies; the study medians vary between .00 and .06; and the mode falls between $\pm$.05!

Why do some schools achieve unexpectedly large reading gains? With so few studies this question cannot answered in any definitive way. There are at most indirect suggestions that such schools may have desegregated in the 1960's, had voluntary plans, included the earlier grades in their evaluation design, been studied for longer time periods, have had a higher percentage of white children in desegregated classrooms, and may have introduced enrichment programs at the same time as desegregation. Such variables could have had independent or joint impacts, and it is inevitable that other variables could be thought of that should be added to any list of possible explanations of why some districts gained so much more than others in reading. Among the possibilities is chance, for it is noteworthy that the outlier studies had smaller sample sizes and that, with the exception of Anderson, the districts with the largest gains in reading

were not the districts with the largest gains in math. While it is not necessary for desegregation to impact on both--and Stephan gives an ex post facto rationale for why desegregation should affect reading but not math--we would be more confident of having identified valid outliers had there been more of a consistency in gains between reading and math.

If the present analysis had not taken place, there would have been what I interpret to be an impressive consistency of results for reading and math combined. When they defined better studies their own way and combined all measures and grades, both Krol and Crain & Mahard reached comparable mean estimates of .10. (For Crain & Mahard the value is derived from the combined results for their randomized experiments and their two longitudinal designs with black segregated controls.) Using their own preferred set of studies and considering math and reading only, the present panelists arrived at estimates varying around this. Armor obtained .04, Miller .12 and Stephan .14, and Wortman .17 when his two strongest designs were weighted and averaged based on part of his sample of 31 studies. These estimates are generally higher than the values of Krol and Crain & Mahard, but not by much. Indeed, I suspect that few commentators would find much of a difference between a gain of one month and of one and one-half months (.10 versus .15).

The present analyses have muddied these waters by suggesting that the means above are noticeably higher than their corresponding medians or modes and by further suggesting that the choice of a measure of central tendency depends in part on knowledge of the distribution of effect sizes in the population. But with such a small sample, the true distribution cannot be confidently ascertained. For those who accept my analyses, I have substituted a low degree

of certainty about the effects of desegregaton for the higher degree that used to pertain but that depended on distributional assumptions that may be wrong. Social science analyses often increase uncertainty, and this is to be preferred to a premature certainty about something wrong or misleading. However, it is even more preferrable to reduce quickly new sources of identified uncertainty. In the present case, this means examining the distributions obtained by Crain & Mahard (1983) for their better studies to see if they are skewed.

7. A Comparison of the Present Results with Crain & Mahard. Crain & Mahard (1983) insist that the effects of desegregation are best assessed from randomized experiments and from studies where desegregated schooling begins at kindergarten or grade one so that the child has never known segregated schooling. When the randomized experiments and the studies with kindergarten and first grade samples were studied separately, Crain & Mahard obtained estimates of .30 in each case. They therefore interpreted this as the best estimate of the effects of desegregation on the achievement of black children. Such an effect is moderately large by many of the (arbitrary) standards used for assessing the effects of educational interventions, as Walberg's essay in this volume attests. It is certainly a more optimistic value than obtained in the metaanalyses reviewed here. Hence, we will consider the estimates of Crain & Mahard in some detail.

It is clear that their estimates decrease to some extent when we consider medians and modes rather than means. Crain kindly supplied me with the distribution of effect sizes for the seven comparisons involving randomized experiments, with Zdep omitted. The mean was .27, the median .24, and the mode

could not be computed. For the kindergarten and first grade samples evaluated using before-after designs and black segregated control groups, the mean based on 17 comparisons was .31, and the median and mode were each .26. I do not know what the mean, median and mode were for all the studies and all the grades with before-after measures and black controls. Nonetheless, the data above suggest that the medians and modes do not reduce to zero in the studies that Crain and Mahard prefer for estimating the effects of desegregation.

Unfortunately, the results of Crain & Mahard are not easy to interpret as estimates of generalized causal impact. First, nearly all the randomized experiments were part of Project Concern and so offer little comfort as to the generalizability of effects. Also, with so few degrees of freedom in the analysis of randomized experiments, it is not likely that the mean effect reliably differs from zero. Second, only one of the kindergarten and first grade samples of Crain & Mahard was included in the present panel's sample--Carrigan--despite the specification of both Crain & Mahard and the present panel that before-after designs and black controls characterized better studies. This discrepancy in the number of comparisons presumably occurs because of differences in strategies used to estimate standard deviations and--principally--because Crain & Mahard were willing to accept pretest measures that the present panel would not accept because it required that pretest and posttest measures tap into the same conceptual domain. For understandable reasons the pretest measures of very young children tend to reflect "academic readiness" rather than the academic achievement that is assessed at the posttest. If the usual selection bias operated and the children attending desegregated schools were more able or more motivated than their segregated

counterparts, then the reduced pretest-posttest correlation caused by differences between the readiness and achievement measures would probably result in overestimating the effects of desegregation in each study (Campbell & Boruch, 1975). Consequently, it is unlikely that valid estimates of the effects of desegregation were obtained with the kindergarten and first grade samples of Crain & Mahard, though the authors have indeed identified a significant issue. After the first generation of desegregation in a district, no students enter desegregated schools from segregated ones--nearly all begin and end their schooling in desegregated classes. Consequently, it is of special importance to learn how desegregation is related to the achievement of very young children.

The estimate of Crain & Mahard that most closely approximates the work of the present panel is based on all grade levels, all outcome measures, before-after designs, and black control groups. As mentioned earlier, the estimate they obtained was .10, and this is much closer to the panel's estimate than the probably inflated value of .30 provided by studies of kindergarten and first grade children where initial differences were not well controlled for. However, nothing in the present panel's work specifically refutes an implicit claim--in Crain & Mahard--that desegregation may have larger impacts at younger grades. To say that .30 may be inflated is not to say the true value for the youngest children is .10! The issue of grade differences in effect sizes has not been solved by either the present panel or Crain & Mahard, and must remain on issue for further research.

INTERPRETATION

I want now to interpret the meaning of both the absence of gains in mathematics and the presence of reading gains of between two and six weeks. To do this, I broach two issues. First, I ask what implications the findings have for various stakeholder groups, and in so doing I also explore how generalizable the findings are beyond the 19 studies examined. Second, I ask what implications this metaanalysis project has for theories of research synthesis.

1. Stakeholder Analysis

(a) Protagonists of School Desegregation. The analyses I have presented might give some comfort to protagonists of school desegregation, particularly those who support it for reasons of equal access, the improvement of race relations, or the enhancement of self-esteem rather than for reasons of academic achievement. For such protagonists the crucial finding from all the analyses of all the scholars is that school desegregation does not decrease the achievement of black children. If it did, this would represent an undesirable side effect of desegregation with which protagonists would probably have to deal ethically, ideologically, and politically. My guess is that it is more difficult to argue that a decrease in achievement is of no consequence than it is to argue that the absence of an increase is of no consequence. Unintentionally decreasing achievement would be a worrisome side effect of desegregation that no protagonist could ignore.

Protagonists of school desegregation can also take some succor from an as yet imperfectly corroborated trend in the data. This is that achievement gains may be larger in younger children who have not had to go through as long a prior experience in segregated classes. Indeed, one of the major points in Crain & Mahard—that we could not independently test—is that achievement gains are greatest of all if black children have never been desegregated. This is a very important point, for many of the advocates of desegregation view it as a means of providing desegregated—or preferably, fully integrated—education to all children for all of their school career. From this perspective, the group of children who start out in segregated schools are not the group of greatest interest. Of more concern are those who have never been segregated and will never experience the historically circumscribed difficulties associated with being among the very first children to transfer into a desegregated school district. Such pioneers move into environments that are novel, not only for them, but also for teachers, administrators, parents and local leaders. Because of the novelty, more mistakes are likely to occur than is the case at a later date when new cohorts of children come through the system, and teachers, administrators and parents should have benefitted from earlier mistakes. Later cohorts might be expected to benefit more from desegregation, both because they have never known segregated schooling and because the school personnel are more experienced with education in mixed racial settings.

Protagonists of desegregation might also note that over half of the studies examined by the present panel involved only one year of desegregation. Moreover, the typical fall-spring testing sessions involve less than a complete school year. Thus, most of the studies involved only a small fraction of the

total time that children experience desegregation, especially if they enter desegregated schools in the early grades. Protagonists of school desegregation might wonder if its full impact has yet been evaluated and they may point to the larger effects in two year studies to suggest that the cumulative impact of desegregation may be much larger than its first year effect. The major problem with this argument is that the studies testing three years of desegregation produced no effects. Consequently, protagonists of desegregation would have to discredit the three-year studies in order to make the case that desegregation has not yet been tested at its presumptively most efficacious. However, it is not difficult to discredit these studies since they are only three in number and they undoubtedly differ from the majority of studies in many ways that are correlated with lower achievement gains.

2. The Perspectives of Antagonists of School Desegregation. The present analyses should bring most succor to antagonists of school desegregation. Where before they would have had to acknowledge the gains in reading caused by desegregation and would have had to argue that their practical implications are trivial--as Armor has done in his present essay--antagonists can now point to analyses which suggest that there have been no real gains in reading because of desegregation in most school districts. This involves a shift in the argument--from how meaningful the obtained reading gains are considered to be, to whether there are any gains at all whose value is worth debating. But although the medians and modes in Tables 1 through 5 could be used by antagonists of school desegregation, I have tried to stress how unstable these estimates are and how much they might be changed by adding just a dozen more

cases to the distribution of effect sizes.

Antagonists of school desegregation can also point to the opaque trend in the data for mandatory programs to result in zero effect sizes and for larger effects to be found with voluntary programs. Few antagonists of desegregation oppose plans in which local authorities agree to desegregate and receiving schools voluntarily accept pupils who volunteer to go to the receiving schools (or whose parents "volunteer" for them). The objection is to mandatory desegregation which, in both my analysis and Stephan's, produced no reading or math gains. (This comparability was achieved despite the fact that Stephan classified only two of the panel's studies as mandatory, whereas using the essays in this volume by Crain and Armor, I classified four as mandatory, although one was by Carrigan.) However, little confidence can be placed in the idea that mandatory desegregation plans cause no reading gains. Given the small number of studies overall, and of mandatory studies in particular, the mandatory/voluntary distinction was correlated with the year desegregation took place, the test used to measure achievement, the region of the country (two studies were in the Dallas/Ft. Worth area), and was probably also correlated with many other factors that would emerge as soon as one examined in detail the specifics of the mandatory desegregation studies by Sheehan & Marcus, Evans, and Van Every.

Antagonists of school desegregation can also point to the paucity of clearcut evidence about desegregation plans that will raise school achievement. Protagonists of school desegregation, and persons whose job it is to plan the desegregation effort in a particular community, want to know what types of desegregation will be effective. They prefer this specific question to the more

global: "How effective is desegregation in general in raising achievement?" All the parties concerned with desegregation research realize that there is no standard desegregation treatment, but many of the protagonists of desegregation hope to discover a set of activities that, when implemented in newly desegregated schools, will raise achievement, among other things. The present analysis has pointed with little confidence to some possible elements of effective desegregation plans. But nothing in the list of elements is new, and after the panel's reviews nothing is better "proven" as a causally efficacious element of desegregation plans than was the case before. Antagonists can point, therefore, to the saliency the present review gives to the continuing uncertainty about the elements of desegregation that enhance achievement. This is not to say that the present metaanalysis probed all—or even most—of the prospective causal elements, or even that it probed the better corroborated among them. All we maintain is that it probed some of them, but failed to make us any more confident that we know how to put together desegregation plans that will raise achievement in reading and math.

(c) Persons Planning Desegregation Activities. Irrespective of their personal beliefs about the desirability of desegregation, mandated or otherwise, there are some groups of persons who have to plan desegregation activities. One such group consists of judges, civil servants, consultants, and school district officials who develop desegregation plans for school districts or metropolitan areas. Such persons want to know about the types of desegregation plan, or the major elements within an overall plan, that will produce the kinds of outcomes they most value from desegregation. The present panel's work provides nothing

of substance to help such planners. It might, however, make a minor contribution to undermining their morale, for the difference in outcomes between the means, medians and modes suggest that the effects of their labors on achievement are likely to be minimal, at least in the short term and to the extent the backward-looking analyses on which this review is based are pertinent to the immediate future.

This last point is crucial. For many theorists of evaluation its function is less to summarize what has happened in the past and more to discover what might be effective in the future. In this context, it is worth noting that the major difficulties with metaanalysis concern the possibility that the bias in one direction may be greater than in the other across all the studies under review. The panelists dealt exhaustively with biases that might lead to false conclusions about whether the relationship between desegregation and learning gains is causal, but few of them considered biases that limit the generalizability of findings and hence their presumed utility for planners. In fact, 16 of the 19 studies were begun in the 1960's, and only one is later than 1975. The dearth of later studies is striking, and Armor's essay contains an important paragraph expressing indignation that so few evaluations of school desegregation were undertaken in the 1970's, a decade characterized by so many large-scale evaluations in other areas within education. Most of the 19 studies under examination were dissertations or local efforts by the staff of a school district. This may explain why the sample sizes are so small, the documentation of desegregation activities so meagre, and the measurement plan so sparse.

Another constant bias is obvious. The panel was constrained to examine how desegregation impacted on the achievement of black children. Yet for most

planners achievement does not exist in a vacuum. The utility of the achievement gains caused by desegregation can vary in meaning depending on whether the desegregation activities in question also reduce or widen achievement gaps between blacks and whites, are or are not accompanied by an increase or reduction in interracial prejudice, are or are not accompanied by white flight, are or are not associated with self-esteem gains, are or are not associated with community support, are or are not related to changes in real estate values, are or are not associated with the founding of magnet or lab schools, etc. By examining just school desegregation and black achievement much of the interpretative context vital to planners is lost.

A second group of planners is composed of teachers, both those contemplating desegregation and those already teaching in desegregated classrooms. In theory, research could be of help to them in identifying practices they can implement that will improve the functioning and results in classrooms. However, the present metaanalytic efforts do not speak to such learning needs. The teacher's needs are more micro than macro, more concerned with process than outcome, and with explanation than descriptive causation. The question on which the panel worked is a question that meets the interests of central government officials with responsibility for oversight more than it meets the interests of those who must plan for desegregation in specific school contexts.

(d) Persons Honestly Seeking to Learn what Desegregation Has Accomplished. The panel's papers help those who would honestly understand what desegregation has accomplished by questioning the utility of so global a label as

"desegregation". Miller's analysis shows that, after the mean effect size is accounted for, more variance remains than is due to chance. This suggests that systematic forces have to be taken into account over and above whether desegregation took place if there is to be any reasonable prediction of effect sizes. Elementary consideration of the decentralized structure of educational decision-making suggests that desegregation plans will differ from location to location and that, even where they appear similar on paper, there will be local adaptations to suit local conditions. From the perspective of someone seeking to learn what desegregation has achieved, elementary questions need to be asked: "What does desegregation mean?"; "What are the criteria that should be used to create clusters of desegregation activities?"; "What types of desegregation result from such clustering procedures?"; and "How well do the different clusters or types of desegregation predict differences in achievement outcomes across districts?". At present, persons interested in learning about school desegregation are more likely to have learned to identify the more pertinent questions than they are to have learned answers to these questions.

But there are some persons interested in the effects of desegregation, very globally conceived, most of whom are government officials with oversight responsibility, journalists, or scholars. The present essay may help sensitize them to the possibility of considerable differences in effects from district to district and to the possibility that, across all districts, effects may be highly variable and even skewed. The possibility of skewness might present them with a problem. Although the mean represents the global impact of desegregation painted on a broad national canvas, it is of no comfort to judges and school districts contemplating desegregation or to teachers worrying about how to

handle a racially mixed class. For some of these people, the mode is more immediately meaningful than the mean. It may be less meaningful in the future, of course, if (1) there really are outliers, (2) the causes of large gains can be explained, and (3) school districts can adopt the causal elements present in the schools with large effects. But we do not yet know what these elements are. In the absence of such knowledge, the differences between the means, medians, and modes highlight anew the conflicting information needs of the many groups in the national educational system who have a stake in desegregation. The differences are most apparent (1) with respect to what should be evaluated--desegregation in general, a specific type of desegregation plan, the particular plan in a particular district, or elements within plans?; and (2) with respect to what should be assessed--achievement, school discipline, race relations, self-esteem, enrollment figures, local tax support for education, local political support for desegregation, home values, etc? But the differences in information needs are also apparent with respect to (3) which measure of central tendency is most appropriate. Different measures speak more to the interests of some stakeholders than others.

2. Theories of Research Sythesis. The present panel represents a unique attempt to probe to what extent experts with three different presumed commitments would converge on a common answer about how desegregation has affected the achievement of black children. Crain and Wortman had already concluded in review articles or papers that desegregation increased achievement; the opposite conclusion has been drawn by Armor and Miller; while Stephan and Walberg had published on the issue but had taken more neutral stances, although Walberg has given court

testimony largely opposed to desegregation. The hope was to achieve a common estimate of effect size despite the different commitments, based on a theory that the results would be more credible, and perhaps even more valid, if they could be replicated across the heterogeneity associated with the analysts' prior professional commitments.

In general, the effect sizes for math and reading combined did reflect the prior commitment. Highest were those of Wortman (.17) and Crain, who stressed the results from his kindergarten and first grade samples and from the randomized experiments he studied (.30 for all outcome measures combined). The next highest estimate was from Stephan (.14 without corrections for length of desegregation), and lowest of all was Armor (.04). The person least fitting expectations was Miller, whose .12 value was intermediate.

Actually, the theoretical rationale for pluralism of analysts was only partially realized, given the decision made before the panel met to restrict the metaanalyses to "good" studies and to use Wortman's prior work to generate that list. One of the major points in metaanalysis where ideology and other commitments enter in is when relevant studies are selected for analysis. Panel members were free to suggest studies for the core list, and Armor succeeded in having two studies added that had negative effect sizes (Sheehan & Marcus, and Walberg). He also made a strong and persistent case for excluding Rentsch and including Carrigan. But few considered calls were heard to add other studies, even though Crain had a list of 93 that he and Mahard considered relevant, more than half of which may have been randomized experiments or longitudinal designs with segregated black control groups. In retrospect, the decision to restrict the selection criteria to a common set rather than let the panelists select

their own, and the failure to assess each of Crain's 93 studies according to the panel's criteria of adequate methodology, may have unnecessarily restricted both the sample of studies and the heterogeneity in assumptions on which the theory behind the use of multiple panelists depends.

It is not difficult to see why the decision was made to restrict the metaanalyses to "better" studies. After all, Krol has found smaller estimates with his "better" studies, as also had Wortman, King and Bryant. But Crain obtained larger estimates with his "better" studies! Obviously, chance differences in the studies available, or differences of opinion about what makes better studies, may have contributed to the apparent puzzle about whether superior methods were associated with larger or smaller effect sizes. Another point is also worth keeping in mind. Although one of the rationales for pluralistic panel members was the credibility and validity afforded by convergence, a second rationale is that divergence in their results might serve to force out the differences in assumptions between advocates and opponents of desegregation, thereby sharpening the focus for future research. Yet the likelihood of such differences being forced out is presumably greater the more freedom panelists have to select studies for review.

Another decision that was made before the panel convened was to use metaanalysis. This technique depends most heavily on the assumption that the average bias is zero with respect to threats to internal, external, construct, statistical conclusion, or any other type of validity (Cook & Leviton, 1980). This assumption is usually dealt with in either or both of two ways. First, a subsample of studies is isolated for which the assumption is made that the bias is zero, and the estimate from this sample is then compared to the estimate for

the remaining subsample where bias might be a problem. If there are no differences in the estimates, the conclusion is drawn that the biasing force in question has not operated. The second strategy is to assume the source of bias away by postulating that the total sample studied is heterogeneous with respect to the threat in question. This last assumption is more credible the more the sample differs on irrelevancies correlated with the major outcomes.

Desegregation research is problematic for the metaanalyst since Wortman has shown that studies without control groups might be biased and few analysts are willing to use norms or white children as "control groups". The need for control groups entails that few studies will meet minimal methodological characteristics. The sample of studies will also tend to be highly variable, given the wide range of desegregation activities in the decentralized education sector and the wide range of children, grades and times studied. Consequently, small samples of possibly abnormally variable estimates will be metaanalyzed. It is difficult to imagine arriving at confident estimates of distribution and central tendencies in this situation; and it is also foolhardy to expect to break the data down in multiple ways so as to examine the correspondence in estimates across different types of desegregation activities, different years when desegregation began, different regions of the country, etc. Consequently, to rule out threats one has to rely on there being "enough" variability in region, year of study, type of activities implemented, etc. But given the small samples, it is not easy to be confident of "enough" heterogeneity in conceptual irrelevancies. Hence, the low level of confidence I have placed in most of my own conclusions and those of the panelists.

These metaanalytic endeavors point to another problem with the method that

overlaps with the problems in using small samples to estimate populations that may be complex and highly variable. Once one has postulated that a skewed distribution may be present, the guiding question becomes the explanatory one: "Why are there outliers?" Explanation is not a strong point of metaanalysis. To explain presumes that we have measures of the potential explanatory constructs for a large sample of studies. Rarely is this the case with metaanalysis, for their availability depends (1) on the extensive measurement of what is implemented as part of a treatment--in the desegregation studies examined, little was available from reports to help with this; and (2) on the extensive measurement of causal micro-mediating processes. For desegregation and reading, such measurement might include, but not be limited to, the assessment of dominant language patterns inside and outside of classrooms. But the sample size of studies with such measures might be expected to be low since the relevant hypothesis about language patterns had not been developed when the earlier evaluators did their work. Indeed, the theory developed <u>because</u> of their work and the anomolies in the data which the work revealed. Since the number of studies with adequate measures of potential explanatory variables will often be low in metaanalysis for reasons of cost and because of the dynamic, evolving nature of theoretical explanatory constructs, metaanalysis will rarely result in confident explanation. This was certainly the case in trying to explain the outliers in Figures 1 through 4. Many potential explanatory forces were isolated, but none of them could be unconfounded from each other with the sample sizes and measures on hand.

CONCLUSIONS

My own reading of the panelists' papers and my own analyses lead me to the following conclusions about how school desegregation has influenced the academic achievement of black students. The conclusions are based on only about 17 studies, and their generalizability is unknown.

1. Desegregation did not cause any decreases in black achievement.

2. On the average, desegregation did not cause an increase in achievement in mathematics.

3. Desegregation increased mean reading levels. The gain reliably differed from zero and was estimated to be between two and six weeks across the studies examined. Only one panelist (Stephan) computed the reading effect per 8 month school year. His estimate is between five and six weeks of gain per year. But since none of the studies involved more than three years of post-desegregation research, it is not possible to compute the mean gain over a child's total school career in desegregated classrooms.

4. The <u>median</u> gains were almost always greater than zero but were lower than the means and did not reliably differ from zero. The <u>modal</u> gains were even less than the median gains and varied around zero.

5. The differences between the means, medians and modes result because the distribution of reading effects appears to be skewed, with a disproportionate

number of school districts seeming to obtain atypically high gains.

6. Studies with the largest reading gains can be tentatively characterized along a number of methodological and substantive dimensions, including: small sample sizes, the study of two or more years of desegregation, desegregated children who outperformed their segregated counterparts even before desegregation began, and desegregation that occurred earlier in time, involved younger students, was voluntary, had larger percentages of whites per school, and was associated with enrichment programs.

7. None of the above factors can be isolated, singly or in combination, as causes of any of the atypically large achievement gains in reading that were obtained in some school districts.

8. The panel examined only 19 studies of desegregation, with most panelists rejecting at least two of them on methodological grounds. When the results for each study (or each comparison) are plotted for reading or mathematics, the distributions are based on so few observations that I could not accept the assumption that the obtained distributions closely approximate what the underlying population distributions are. Because of the small samples and apparently non-normal distributions, little confidence should be placed in any of the mean results presented earlier. I have little confidence that we know much about how desegregation affects reading "on the average" and, across the few studies examined, I find the variability in effect sizes more striking and less well understood than any measure of central tendency.

REFERENCES

Campbell, D. T. and Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate results. In C.A. Bennett & A.A. Lumsdaine (Eds.), Evaluation and experience: Some critical issues in assessing social programs. New York: Academic Press, 1975.

Cook, T. D. and Leviton, L. Reviewing the literature: A comparison of traditional methods with meta-analysis. Journal of Personality, 1980, 48, 449-472.

Crain, R. L. and Mahard, R. E. Desegregation plans that raise black achievement: A review of the research. Santa Monica, CA: Rand Corporation, June 1982.

Crain, R. L. and Mahard, R. E. The effect of research methodology on desegregation achievement studies: A metaanalysis. American Journal of Sociology, 88, 1983.

Glass, E. V. McGaw, B., and Smith, M. L. Meta-analysis in social research. Beverly Hills, CA, Sage Publications, 1981.

Krol, R. A. A metaanalysis of comparative research on the effects of desegregation of academic achievement. Unpublished dissertation, 1978. Ann Arbor, Michigan: University Microfilms (# 6907962), 1979.

Wortman, P. M., King, C., and Bryant, E. B. Metaanalysis of quasi-experiments: School desegregation and black achievement. Ann Arbor, Michigan: Institute for Social Research, 1982.

Table 1

Central Tendencies for Reading - Author's own Preferred Studies

	Mean	Median of Comparisons	Median of Studies	Midpoint of Modal Category of Comparisons
Armor	.06	.00	.00	-.05 & +.05
Miller	.16	--	.06	-.05 & +.05
Stephan	.14	.08	.08	+.05
Wortman[a]	.26	.15	.04	--

[a] In Wortman's case "preferred" studies refers to those of his selection from the panel's core 19 for which pretest adjustments could be made. It does not refer to his analysis of 31 studies.

Table 2

Central Tendencies for Reading - 17 Common Core Studies

	Mean	Median of Comparison	Median of Studies[e]	Midpoint of Modal Category of Comparisons
Armor[a]	.13	.03	0	-.05 & +.05
Miller[b]	.16	--	.06	-.05 & +.05
Stephan[c]	.13	.07	.08	+.05
Wortman[d]	.26	.15	.04	--

[a] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted; Rentsch added and given Wortman values.

[b] Based on N of studies; Carrigan and Thompson & Smidchens omitted.

[c] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted. Thus, Iwanicki & Gable and Slone added.

[d] Based on N of comparisons. The sample size is considerably smaller than with other analysts, since Wortman omitted all instances where the control group standard deviation was not specifically given. This resulted in the omission of Clark, Evans, Iwanicki & Gable, Klein, Lard & Weeks, Slone, Syracuse, and Walberg, as well as Carrigan and Thompson & Smidchens. No mode was ascertainable.

[e] The medians are from Miller's Table 2 for each author based on N of studies rather than comparisons.

Table 3

Central Tendencies for ES Values in Math - Author's own Preferred Studies

	Mean	Median of Comparison	Median of Studies	Midpoint of Modal Category of Comparisons
Armor	.01	-.05	-.06	--
Miller	.08	--	.07	--
Stephan	.04	.02	.02	--
Wortman	.08	-.02	-.05	--

[a] In Wortman's case "preferred" studies refers to those of his selection from the panel's core 19 for which pretest adjustments could be made. It does not refer to his analysis of 31 studies.

Table 4

Central Tendencies for Reading and Math Combined - Authors' own Preferred Studies

	Mean	Median of Comparisons	Median of Studies	Midpoint of Modal Category of Comparisons
Armor	.06	.00	.00	-.05
Miller	.12	--	.06	-.15 & +.05
Stephan[b]	.07	.05	.05	-.05
Wortman[a]	.16	.08	.01	-.05

[a] In Wortman's case "preferred" studies refers to those of his selection from the panel's core 19 for which pretest adjustments could be made. It does not refer to his analysis of 31 studies.

[b] These are estimates per school year.

Table 5

Central Tendencies for Reading and Math - 17 Common Core Studies

	Mean	Median of Comparisons	Median of Studies[e]	Midpoint of Modal Category of Comparisons
Armor[a]	.08	0	0	-.05
Miller[b]	.12	--	.06	-.15 & +.05
Stephan[c]	.07	.03	.06	+.05
Wortman[d]	.16	.08	.01	-.05

[a] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted; Rentsch added and given Wortman values.

[b] Based on N of studies; Carrigan and Thompson & Smidchens omitted.

[c] Based on N of comparisons; Carrigan and Thompson & Smidchens omitted. Thus, Iwanicki & Gable and Slone added. Estimates of effect per school year.

[d] Based on N of comparisons. The sample size is considerably smaller than with other analysts, since Wortman omitted all instances where the control group standard deviation was not specifically given. This resulted in the omission Clark, Evans, Iwanicki & Gable, Klein, Laird & Weeks, Slone, Syracuse, and Walberg, as well as Carrigan and Thompson & Smidchens.

[e] The medians are from Miller's Table 2 for each author based on N of studies rather than comparisons.

Figure 1: Distribution of Reading Effect Sizes in Armor

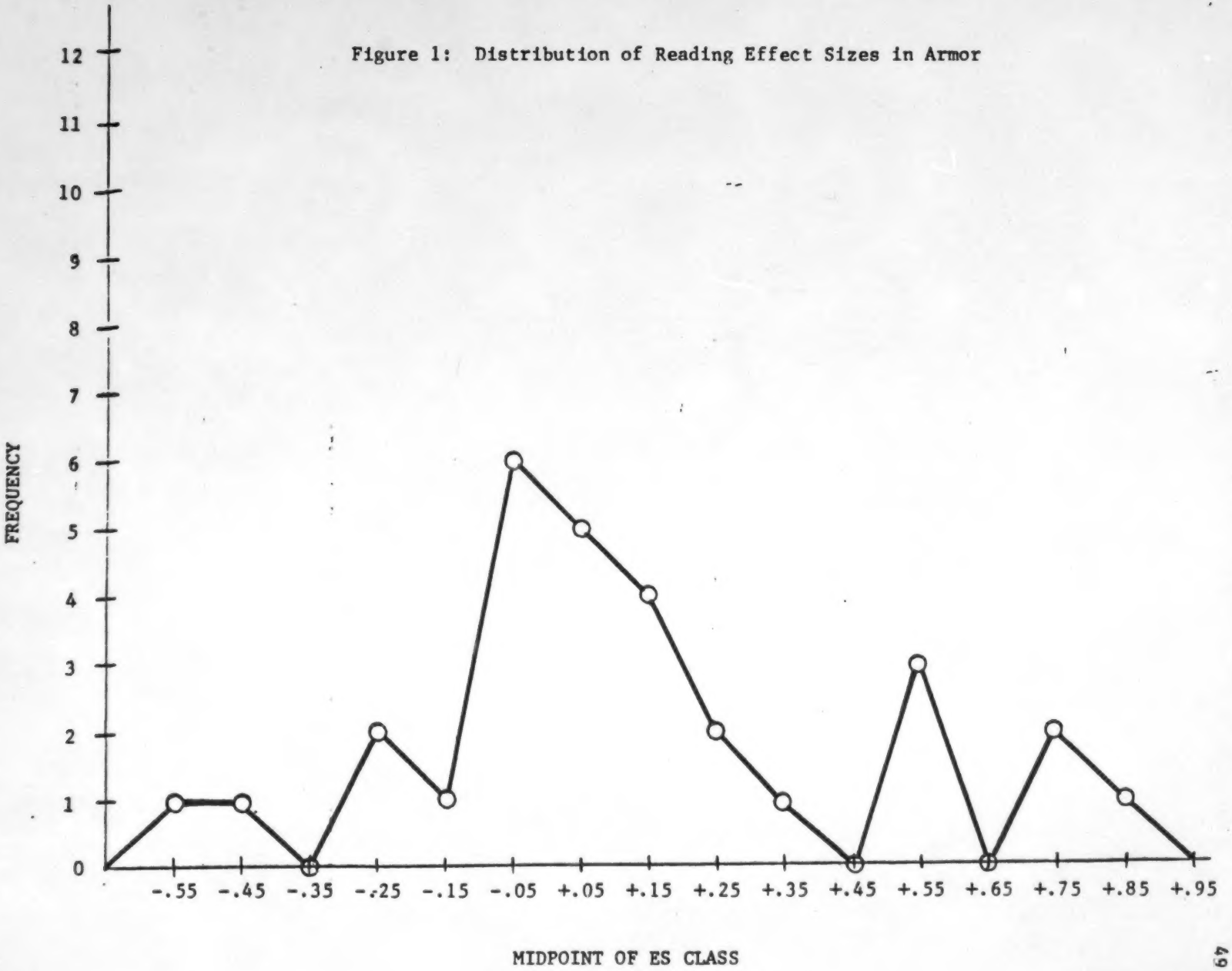


Figure 2: Distribution of Reading Effect Sizes in Miller

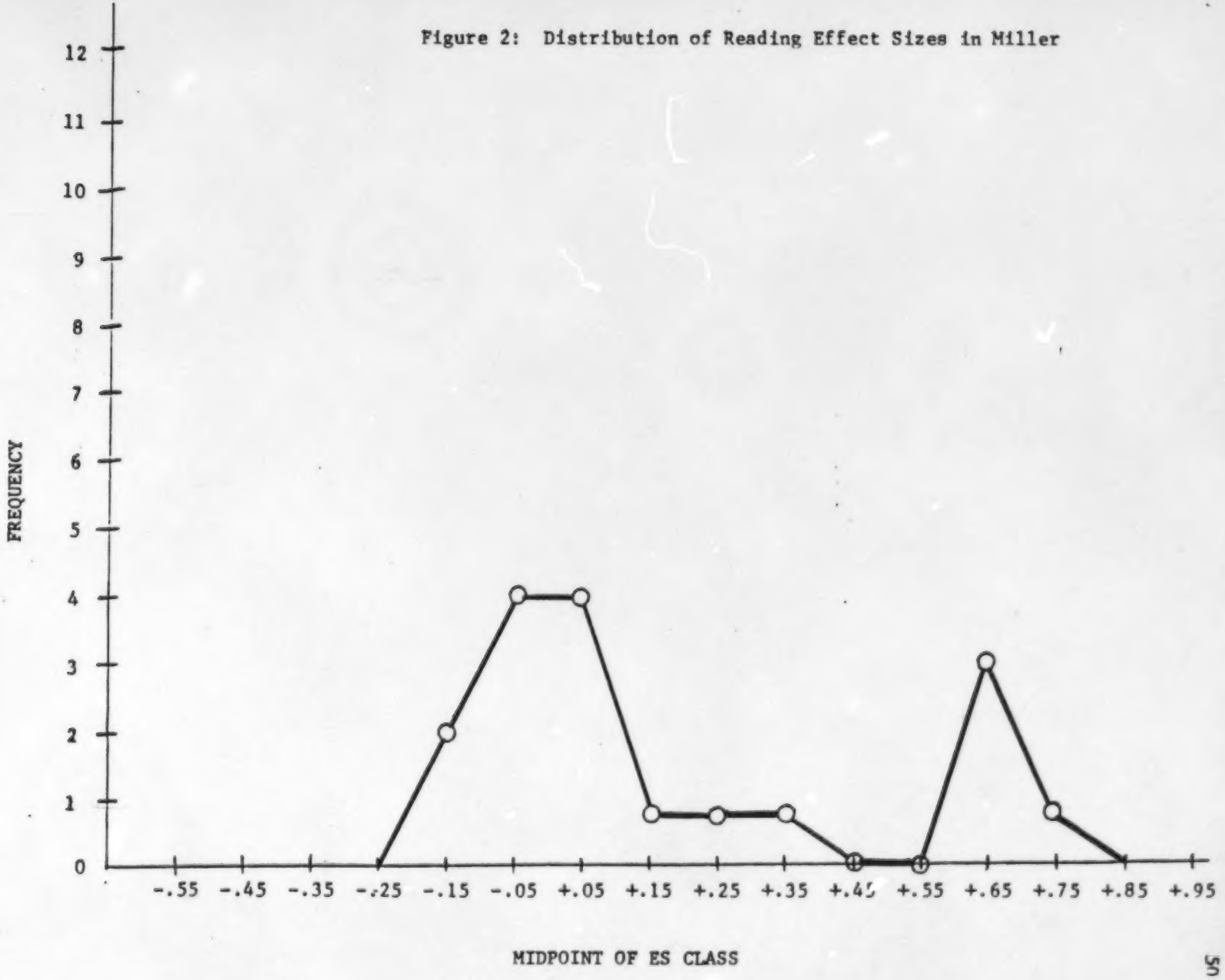


Figure 3: Distribution of Reading Effect Sizes in Stephan

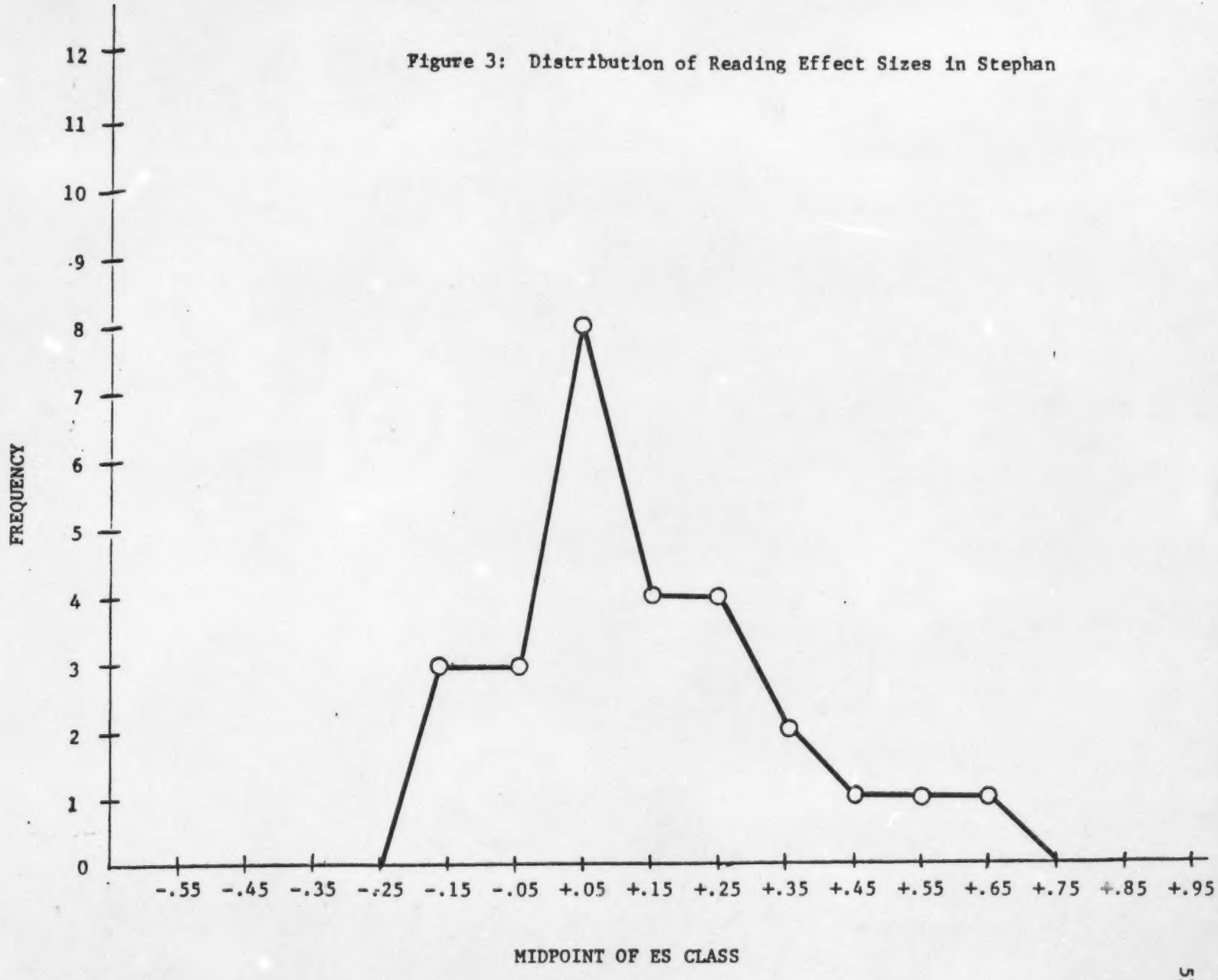


Figure 4: Distribution of Reading and Math Effect Sizes Combined for the Pretest-Adjusted Studies of Wortman

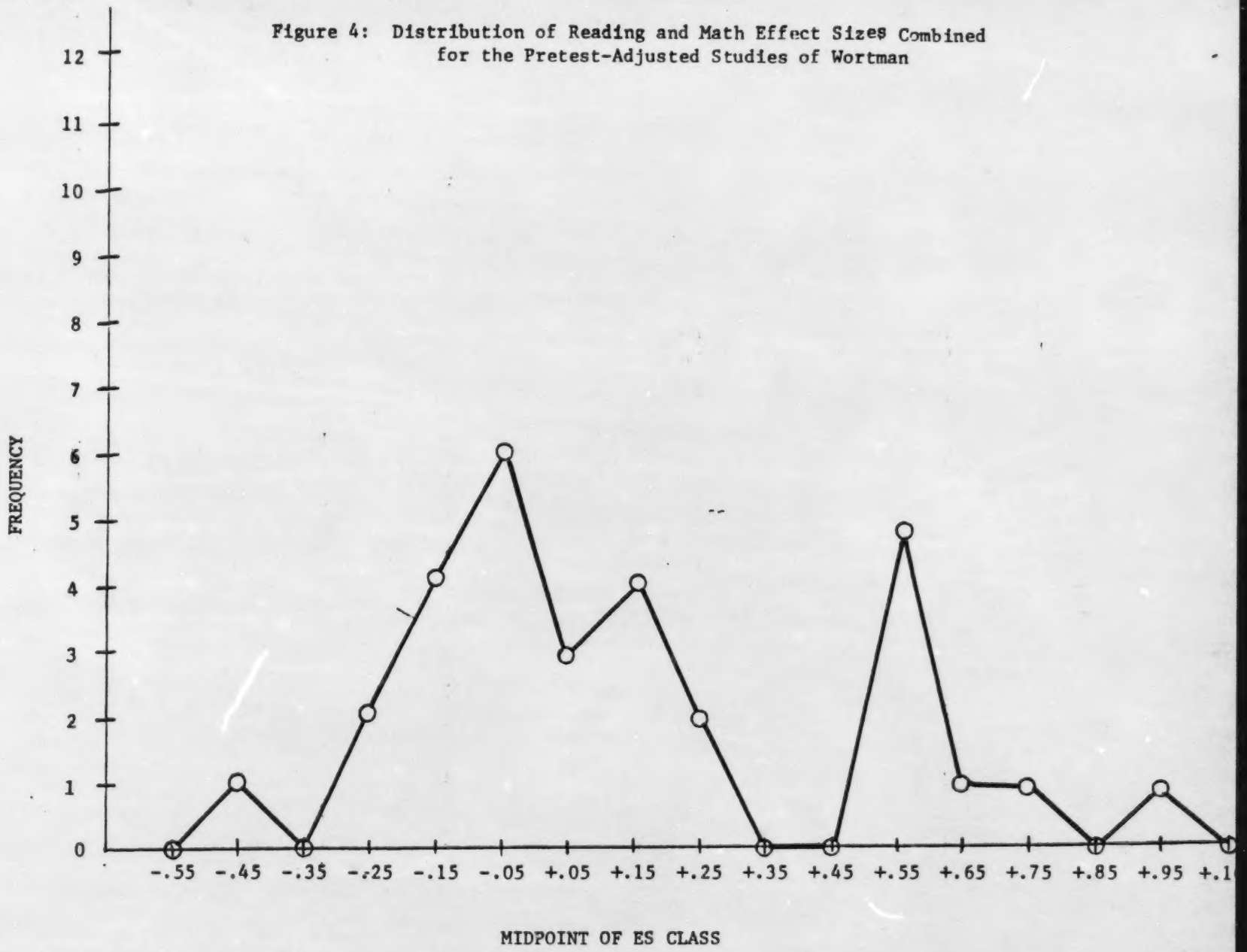
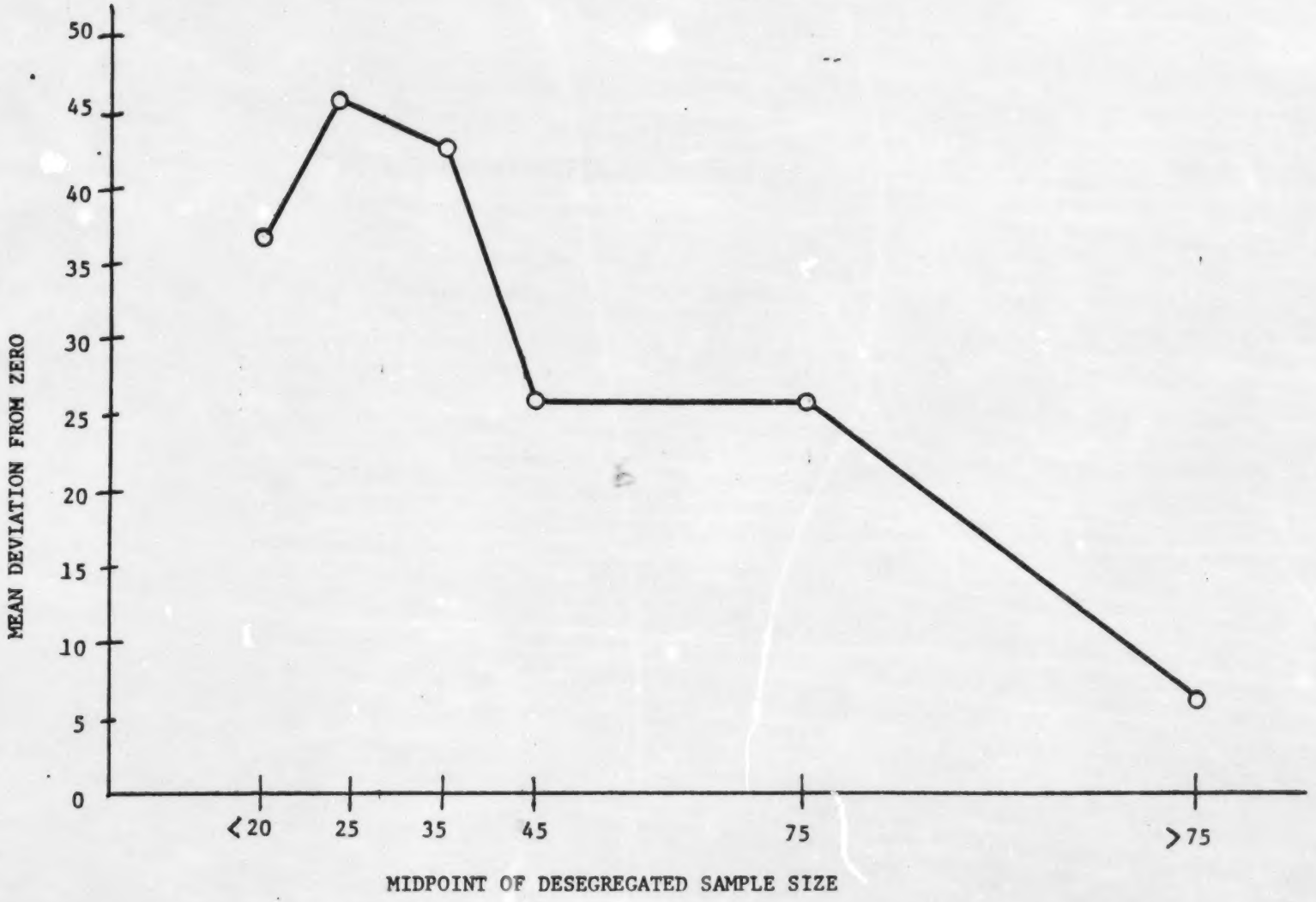


Figure 5: Relationship between Sample Size and Magnitude of Effect Size Irrespective ro their Sign



Reading

Figure 6: Relationship between Grade Level at Desegregation and Mean Effect Size per Eight Months of Desegregation